Word Spotting in Background Music

a Behavioural Study

Marchegiani, Letizia; Fafoutis, Xenofon

Published in:
Cognitive Computation

DOI (link to publication from Publisher):
10.1007/s12559-019-09649-9

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Cognitive Computation manuscript No.
(will be inserted by the editor)

Word Spotting in Background Music: A Behavioural Study ⋆

Letizia Marchegiani · Xenofon Fafoutis

Received: date / Accepted: date

⋆ This is a post-peer-review, pre-copyedit version of an article published in Cognitive Computation. The final authenticated version is available online at: http://dx.doi.org/10.1007/s12559-019-09649-9

Letizia Marchegiani, Corresponding Author
Department of Electronic Systems
Aalborg University
Fredrik Bajers Vej 7, Aalborg Ø, 9220, DK.
E-mail: lm@es.aau.dk

Part of this work was conducted while Letizia Marchegiani was with the University of the Basque Country and the University of Oxford.

Xenofon Fafoutis
Department of Applied Mathematics and Computer Science,
Technical University of Denmark,
2800 Kgs. Lyngby, Denmark.
E-mail: xefa@dtu.dk

Structured Abstract

Introduction

Speech intelligibility in realistic environments is directly correlated with the ability of focusing attention on the sounds of interest while discarding the background noise and other competing stimuli. This work investigates task-driven auditory attention in noisy environments. Specifically, this study focuses on the ability to successfully execute a word spotting task while speech perception has to cope with by the presence of music playing in the background.

Methods

The executed behavioural experiments consider different types of songs and explore how their distinct characteristics (such as dynamics or presence of distortion sound effects) affect the subjects' task performance and, thus, the distribution of attention.

Results

Our results show that the ability of correctly separating the target sound from the background noise has a major impact on the performance of the subjects. Indeed, songs not presenting any distortion effect result to be more distracting than the ones with distortion, whose frequency spectrum envelop differentiates more from the one of the narrative. Furthermore, subjects performed the worst with songs characterised by high dynamics playing in the background, due the unexpected changes capturing the attention of the listener.

Keywords

Keywords Speech Perception · Auditory Attention · Word Spotting · Cocktail Party · Auditory Masking · Music Perception

PACS *43.71.-k, 43.71.+m · *43.72.-p, 43.72.+q · *43.75.Zz

1 Introduction

In everyday life, speech perception is constantly challenged by the presence of different kinds of overlapping noise. Several factors concur to reduce, at various levels, both speech intelligibility and the listeners' performance at semantically analysing the incoming stimuli. For a review, see [1]. For an illustration of how acoustic information is hierarchically processed in the auditory cortex, see [2]. The correlation between visual and auditory elements in speech perception has been also investigated (see [3] for a detailed review), as well as potential benefits of multi-modal settings in hearing aids technology [4]. Some of these factors affecting speech intelligibility are directly correlated to physical constraints, such as signal degradation, and energetic and modulation masking [5]. Some others are attributed to *informational masking* [6], in the form of cognitive load, as well as attention filtering [7], and proficiency in the language [8, 9]. In either cases, the ability of the listener to isolate the sound of interest from the background noise represents a *conditio sine qua non* to avert selective attention failures. In 1953, Colin Cherry [10] introduced the term *Cocktail Party* effect to indicate the ability of humans, in presence of several signals, to focus on specific stimuli while discarding others, and investigated potential causes influencing such ability. Several experiments confirmed Cherry's initial hypotheses, suggesting that some specific factors could actually help the mental ability of performing this segregation and extracting one specific signal from the environment, such as differences in the features of the competitive stimuli [11], the presence of a task [12], and the spatial location of the involved sound sources [13, 14].

This paper focuses on task-driven attention in a simulated noisy scenario, in which speech perception has to cope with the presence of background music. Indeed, the attentive mechanisms in background music are interesting in several applications that range from scenarios where attention is critical for safety (e.g. driving or operating machinery) to social scenarios (e.g. organisation of social events). Furthermore, the analysis of such mechanisms can provide useful insights for fine tuning of hearing aids in similar circumstances.

Parente [15] examined the correlation between music preference and music distraction. A more general analysis of this correlation in a marketing scenario has been proposed in [16]. Later, North et al. [17] extended this analysis, exploring also the distracting efficacy of either arousing or relatively unarousing music. Other related works investigated the impact of loudness [18], and the effect of music tempo on reading abilities [19]. The effects of auditory selective attention on the processing of syntactic information in music and speech has been analysed by Maidhof et al. [20], while the effect of music knowledge on speech perception is illustrated by Slater et al. [21]. A detailed review on the effects of music on verbal learning and memory is provided in [22].

This work presents an analysis of task-driven attention in a simulated cocktail party scenario, in which the voice of interest is masked by alternating songs with different characteristics. Our analysis complements the above mentioned literature, investigating the distracting effect of specific kinds of music on the

performance in a word spotting task. Specifically, with respect to previous works (*e.g.* [18, 19]) which focus on tasks involving reading and/or computational efforts, we explore how the music's dynamics, as well as the presence of distortion effects influence speech perception and intelligibility. In this perspective, this investigation shares the same aspirations of [12–14] to evaluate the impact of stimulus physical characteristics on stream segregation and distribution of attention.

Previous works (e.g. [23]) on visual attention explored the effect of the similarity between target and distractor on object identification. More recently, [24, 25] investigated the concept of *motion saliency* exploiting information on objects' dynamics to model the distribution of attention. In this paper, we perform a similar analysis in an auditory context, by evaluating the impact of both static aspects (in terms of frequency spectrum similarities between the target and the distractors) and dynamics variations (in terms of temporal changes in the frequency spectrum) of different types of distractors. In [11], the authors carry out analogous experiments, focusing rather on speech identification and comprehension performance, without taking into account attentive aspects. In this work, instead, we focus on the attentive elements involved and, inspired by the concept of *motion saliency*, expressed in [24, 25], we aim to analyse the role played by music dynamics in the distribution of attention.

More specifically, in order to understand how certain properties affect speech intelligibility and the performance in a word spotting task, a series of behavioural experiments are carried out, asking the subjects to follow a narrative and push a button every time they hear a specific word. The presented investigation is twofold. First, the influence of energetic masking (in the form of temporal and spectral overlap) applied by the background music is explored. The aim is to exclude the case that the subjects' performance is due to their inability to hear the voice of the speaker. Having excluded that, the influence of the songs on the distribution of attention and speech perception is analysed. To the best of our knowledge, this is the first work investigating the effect of music dynamics and distortion effects on distribution of auditory attention and speech intelligibility. We believe that the use of rock music together with the choice of relying on a word spotting task, provide a valuable framework for subjects' evaluation in a realistic everyday scenario.

2 Methods

The experiment was constructed as follows. The participants were asked to focus their attention on a narrator and identify a specific word by clicking on a button, while different songs were alternating in the background. The narrative was a fairy tale [26]. Targeted for children, the fairy tale uses simple language that is easily understandable by non-native English speakers. The subjects were asked to identify the word *'and'*. Since the selected word is very common and it can be easily missed, the participants' full attention is required to successfully perform the task. Furthermore, with such a common

word bottom-up cues that depend on the rarity of the sound and the possible 'surprise effect' [27] are avoided. During the first 2 minutes of the narrative, there was no background music. During the remaining 6 minutes, three songs were alternating in the background. Each song played for 2 minutes. The target word, *'and'*, appears 9 or 10 times in each 2-minute time slot, resulting to a total amount of 38 word appearances. The average and the minimum time-distance between two occurrences of the target word was 12.3 s and 0.7 s respectively.

We chose to use rock music as masking background because it is shown to be particularly distracting from performing a task [28]. The three selected songs are characterised by the existence of distortion sound effects (or lack thereof) and by high dynamics, that is songs that frequently alternate between clean states and aggressive states with distortion sound effects through the duration of the song.

The experimental work carried out by [29] indicates that the subjects' familiarity to the songs can influence the distribution of attention in different ways, making it particular hard to generalise their effect, especially when the song becomes an emotional trigger. Indeed, previous studies [30] investigated the neural correlates of human emotional judgement, suggesting a strong correlation between attentional mechanisms and emotion analysis. To mitigate any influence of the subjects' familiarity to the background music, unpopular songs were selected based on the statistics from *Last.fm* music social network. In particular, the songs were selected among songs that have less than $350K$ unique listeners worldwide. The subjects' unfamiliarity to the chosen songs was later verified with post-questionnaires. Table 1 summarises the selected songs, which for the remainder of the paper will be referred to as ND (No Distortion), D (Distortion) and HD (High Dynamics). When mixed with the narrative, the Root Mean Square (RMS) energy of the songs was normalised to the same level, which was sufficiently low to ensure that there is no overload distortion after mixing. The transition between two consecutive songs was smoothed out using fading. Moreover, the energy of the narrative was adjusted to 3 dB higher than the background music (*i.e.* Signal-to-Noise Ratio, $\texttt{SNR} = 3$ dB). The SNR level was chosen empirically after a series of pilot experiments, so that the narrative could be clear and intelligible.

The songs were mixed in two different orders between which, the subjects were divided. The purpose of this is to mitigate the influence of the subjects' fatigue on the results. The sounds were presented to the participants remotely via a web-interface, similarly to [11]. The participants were instructed to use headphones and perform the experiments in silent environments. The post-questionnaire confirmed that they adhered to the instructions. Prior to the actual experiment, the subjects were asked to do a 1-minute test experiment to get familiar with their task and adjust the playback level. A total amount of 22 subjects (between $25-35$ years old), all non-native, yet fluent English speakers, with no hearing, or language impairment, participated in the experiment (11 subjects per song order). Their task performance, their answers to the post-

questionnaire and occasional short interviews suggest that all the subjects understood their task at an adequate level.

3 Results and Discussion

Our analysis is articulated in two different steps. We first explore the impact of energetic masking (in the form of spectral and temporal overlap) on the subjects' performance and then investigate whether and how this performance is affected by the songs' characteristics. The subjects' performance is here expressed as successful identification of the word *'and'*. Specifically, we consider as true positives any word identification that occurred between 0.1 seconds and 2.5 seconds from the actual word appearance in the narrative. All other word identifications are considered false positives. Figure 1 shows the statistical distribution of the precision and the F1 score for each song, over the 22 subjects. The mean and standard deviation are provided in Table 2. The relatively high performance in the absence of background music shows that the subjects were able to perform the task. The analysis of the results continues as follows. First, the correlation between the subjects' performance and the temporal and spectral overlap of the narrative and the background music is examined. Assuming that such correlation does not exist, the relative performance variation in presence of various background music can be attributed to the properties of the songs.

3.1 Spectral and Temporal Overlap

The spectral and temporal overlaps, introduced by the musical background, are computed using of the concept of Ideal Binary Mask (IBM). As highlighted by Wang [31], IBMs are defined according to the nature of the signal of interest and their performance is similar to the way the human auditory system functions in the presence of masking. Further investigations [32] have shown that these masks can be exploited to improve the speech reception threshold and, more generally, speech intelligibility. IBMs have also been used to calculate the masking between two narratives uttered by a speech synthesiser in a monaural combination [12]. The same approach is applied in this work to estimate spectral and temporal overlaps between the story and the background music, and their relative effect on speech intelligibility. A binary mask is a binary matrix in which 1 represents the most powerful time-frequency regions of the target signal compared to an interference signal, according to a local criterion (LC). If $T(t, f)$ and $I(t, f)$ denote the target and interference time-frequency magnitude, then the IBM is defined by the following formula.

$$IBM(t, f) = \begin{cases} 1, & \text{if } T(t, f) - I(t, f) > \texttt{LC}. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The time-frequency representation is based on the model of the human cochlea, by the use of gammatone filtering [33]. The parameters controlling the structure of the binary masks are the LC, the window length (WL) and the number of frequency channels (FC). The masking between each audio frame containing the word *'and'* in the story and the respective frame in the song sequence is estimated. This estimation is based on the comparison between the IBMs correspondent to each pair of frames [12, 29]. The spectral overlap is determined by the co-occurrence of 1 in the two binary masks over the total number of time-frequency bins. The temporal overlap is obtained by compressing the IBMs over frequency, assigning value 1 if there is at least a single 1 in one of the relative frequency bins and 0 otherwise. The resulting binary vectors, which we name Compressed Ideal Binary Masks (CIBM), are compared and the amount of temporal overlap is given again by the number of co-occurrence of 1 on the CIBMs over the total number of bins in the vectors. The parameters of the IBMs (LC, WL and FC) are optimized to maximise the correlation. Even with optimised parameters, the correlation is very weak (spectral correlation is 0.14 for the first song order and 0.17 for the second song order; temporal correlation is 0.04 for both). A permutation test with 10000 resamples at 5% significance level did not show a significant correlation ($p > 0.11$). Without significant correlation between the masking level and the ability of the subjects to identify the requested words, we mainly attribute the difference in the performance of the subjects to the song properties.

3.2 Analysis of Song Influence

The results shown in Figure 1 suggest a significant difference in the performance (both in terms of precision and F1 score) when no music is present and when the various songs are alternating in the background. A repeated-measures ANOVA (Analysis Of Variance) with masker type as factor confirmed the statistical significance of this difference both in terms of precision ($\mathtt{F}(3, 84) = 16.43,\ \mathtt{p} < 0.01$), and F1 score ($\mathtt{F}(3, 84) = 14.72,\ \mathtt{p} < 0.01$). The Mauchly test verified that the sphericity condition was not being violated in either cases. The song with distortion sound effects (D) results to be the least effective masker, allowing better performance for the subjects, while the song characterised by high dynamics (HD) appears to be the most distracting. Post hoc tests (with Bonferroni adjustment for multiple comparisons) proved a significant difference ($p < 0.01$) in the performance of the subjects between the D-ND and the D-HD maskers. No significant difference in the performance have been observed between the HD-ND maskers. In case of the former, the significantly higher performance is attributed to the fact that distortion sound effects introduce high frequency components in the music, which are easier to separate from the narrator's voice. Thus, auditory segregation and following grouping are simpler. Indeed, the greater the difference between the features of two sounds, the easier the segregation process [10]. To corroborate the intuition, we investigate the perceptual separability between the D, the ND songs

and the narrative. Following traditional approaches [34], we evaluate the similarity between the D and ND songs and the correspondent sections in the narrative, employing Mel-Frequency Cepstrum Coefficients (MFCCs) [35] to represent the spectral envelopes of the signals. MFCCs mimic the logarithmic perception of the magnitude and pitch of the human auditory system, and provide a representation of the short-term power spectrum of a sound, based on a linear cosine transform of the log power spectrum on a non-linear Mel scale of frequency [35]. In this work, we use the first 12 coefficients plus the relative delta and delta-delta ones. The delta and delta-delta coefficients, also called *velocity and acceleration coefficients*, are commonly used to describe the dynamics of a signal, expressed as the evolution of the MFCCs over time. The *delta coefficients* $\Delta(t)$, which refer to the first derivatives of the MFCCs $x(t)$ are obtained by

$$\Delta(t) = \frac{\sum_{n=1}^{N} n(x(t+n) - x(t-n))}{2\sum_{n=1}^{N} n^2}, with\ N = 2 \tag{2}$$

The *delta-delta coefficients* $\Delta\Delta(t)$, which refers to the second derivatives of the cepstrum coefficients are then obtained in a similar fashion from the *delta coefficients*. The Radial Basis Function (RBF) is computed to convert distances in the MFCC space to a dissimilarity measure. The analysis is performed by dividing the signals in frames of 1s with a sliding overlapping window of 100 ms. The ND song exhibits a higher similarity score (i.e. harder perceptual separability) to the relative narrative counterpart, compared to the song where distortion is present. More specifically, the average dissimilarity between the D song frames and the narrative ones is 12544, while the dissimilarity between the ND song frames and the narrative ones is 8410, yielding a relative difference greater than 60%. A repeated-measures ANOVA with masker type as factor confirmed the statistical significance of this difference ($\mathtt{F}(1, 238) = 17.75,\ \mathtt{p} < 0.01$). The Mauchly test verified that the sphericity condition was not being violated. Since the subjects are unfamiliar with the song, in case of songs with high dynamics, the frequent and sudden changes in the song's dynamics are unexpected, and therefore distracting [27]. To evaluate those changes and validate our results, we consider the difference in the spectral envolope between consecutive frames. Also in this case, we employ MFCCs to represent the envelopes and RBF to compute the distances between those. In this case, as we are taking time explicitly into account in the computation, we consider only the first 12 coefficients, discarding the relative first and second derivatives. The analysis is carried out by first dividing the signals in frames of 1s with a sliding overlapping window of 100 ms and the calculating the distances over consecutive 100 ms windows. The total change for each frame is computed as the sum of the differences across all the windows. A repeated-measures ANOVA with masker type as factor confirmed the statistical significance of this difference ($\mathtt{F}(1, 238) = 534.76,\ \mathtt{p} < 0.01$). The Mauchly test verified that the sphericity condition was not being violated. Figure 2 shows the MFCCs for a sample frame from the narrative and from the three different song types.

Finally, the effect of the different maskers on the response time of the subjects is investigated. Figure 3 shows the statistical distribution of the response time, for all subjects per different masking conditions. The mean and standard deviation are provided in Table 3. A repeated-measures ANOVA with masker type as factor confirmed a statistical significance in the response time ($\mathtt{F}(3, 84) = 19.47,\ \mathtt{p} < 0.01$) when no music is present and when the various songs are alternating in the background. The Mauchly test verified that the sphericity condition was not being violated. Yet, post hoc test did not show significant difference in the response time between the three different maskers.

4 Conclusion

This work presents a behavioural experiment that investigates task-driven attention in a simulated cocktail party scenario, characterised by the presence of background music. The subjects were asked to perform a word spotting task while different songs were alternating in the background. In line with previous studies in different acoustic scenarios and masking conditions, our results show that the ability of correctly separating the target sound from the background noise has a major impact on the performance of the subjects. Indeed, songs which did not include any distortion effect proved to be more distracting than the ones with distortion, whose frequency spectrum envelope differentiates more from the one of the narrative. In addition, subjects performed the worst with songs characterised by high dynamics playing in the background, due to the unexpected changes capturing the attention of the listener. These findings confirm analogous results in the visual domain [27], highlighting the resemblance in the behaviour of the two attention modalities. The authors had no direct control on experimental conditions such as the type of headphones used, the exact noise level in the environment, or the Sound Pressure Level (SPL) generated at the eardrum. Such uncontrolled framework might have added some noise to the results. However, given that the subjects were provided with detailed instructions on how to perform the experiments, including notes on the environmental setting (*e.g.* silent environment, use of headphones), and post-questionnaires confirmed that such instructions were, indeed, followed, we believe the impact of the above mentioned uncontrolled conditions, if any, was only limited. Furthermore, the nature of this setup allows the evaluation of the subjects' performance in a realistic everyday setting, and might be useful to drive the development of specific acoustic technologies in applications that range from scenarios where attention is critical for safety to hearing aids design. Previous works [17, 29] proved that the emotional state induced in the subject by a stimulus can highly influence the distribution of attention. Other investigations [36] provided an analysis of different music emotion recognition approaches, in the attempt of linking specific features of songs to the emotional state they could trigger. The exploration of the emotional character of the songs (considering both arousal and valence effects) and their influence on word spotting and task-driven auditory attention against background music

are directions for future work. Future work may also consider other types of background music, such as classical music [37]. Ultimately, we plan to further investigate the effect of music variability across frequency and its potential impact on its predictability and distribution of attention.

Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

1. Sara Guediche, Sheila Blumstein, Julie Fiez, and Lori L Holt. Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. *Frontiers in Systems Neuroscience*, 7.126, 41–56. (2014).

2. Kazuhisa Fujita, Yusuke Hara, Youichi Suzukawa, and Yoshiki Kashimori. Decoding word information from spatiotemporal activity of sensory neurons. *Cognitive Computation*, 6(2):145–157. (2014).

3. Andrew Abel and Amir Hussain. Novel two-stage audiovisual speech filtering in noisy environments. *Cognitive Computation*, 6(2):200–217. (2014).

4. Amir Hussain, Jon Barker, Ricard Marxer, Ahsan Adeel, William Whitmer, Roger Watt, and Peter Derleth. Towards multi-modal hearing aid design and evaluation in realistic audio-visual settings: Challenges and opportunities. *Proceedings of the 1st International Workshop on Challenges in Hearing Assistive Technology*. (2017).

5. Michael A. Stone, Christian Füllgrabe, and Robert C Mackinnon, and Brian CJ Moore. The importance for speech intelligibility of random fluctuations in steady background noise. *The Journal of the Acoustical Society of America*, 130.5, 2874–2881. (2011).

6. Gerals Jr Kidd, Christine R Mason, Virginia M Richards, and Nathaniel I Durlach. Informational masking. *Auditory perception of sound sources*, 143–189, Springer. (2008).

7. Holger Mitterer and Sven L Mattys. How does cognitive load influence speech perception? an encoding hypothesis. *Attention, Perception, & Psychophysics*, 79(1):344–351. (2017).

8. Martin Cooke, ML Garcia Lecumberri and Jon Barker. The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, 123.1, 414–427. (2008).

9. Letizia Marchegiani and Xenofon Fafoutis. On cross-language consonant identification in second language noise. *The Journal of the Acoustical Society of America*, 138.4, 2206–2209. (2015).

10. E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, Melville, NY, USA. (1953).

11. L. Marchegiani, X. Fafoutis, and S. Abbaspour. Speech Identification and Comprehension in the Urban Soundscape. *Environments*, 5(5), 56. (2018).

12. Letizia Marchegiani, Seliz G Karadogan, Tobias Andersen, Jan Larsen, and Lars Kai Hansen. The role of top-down attention in the cocktail party: Revisiting cherry's experiment after sixty years. In *Proceedings of the IEEE International Conference on Machine Learning and Applications and Workshops (ICMLA)*, 1:183–188. IEEE, New York, NY, USA. (2011).

13. Edward J Golob, K Brent Venable, Jaelle Scheuerman, and Maxwell T Anderson. Computational modeling of auditory spatial attention. In *Annu. Conf. Cogn. Sci. Soc*, volume 39. (2017).

14. Jacques A Grange and John F Culling. The effect of listener head orientation on speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 141(5):3971–3971. (2017)

15. JA Parente. Music preference as a factor of music distraction. *Perceptual and motor skills*, 43(1):337–338, SAGE Publications, New York, NY, USA. (1976).

16. Zohreh Gholami Doborjeh, Maryam G Doborjeh, and Nikola Kasabov. Attentional bias pattern recognition in spiking neural networks from spatio-temporal eeg data. *Cognitive Computation*, 10(1):35–48. (2018)

17. Adrian C North and David J Hargreaves. Music and driving game performance. *Scandinavian Journal of Psychology*, 40(4):285–292, John Wiley & Sons, Haboken, NJ, USA. (1999).

18. David E Wolfe. Effects of music loudness on task performance and self-report of college-aged students. *Journal of Research in Music Education*, 31(3):191–201, SAGE Publications, New York, NY, USA. (1983).

19. Kari Kallinen. Reading news from a pocket computer in a distracting environment: effects of the tempo of background music. *Computers in Human Behavior*, 18(5):537–551, Elsevier, Amsterdam, Netherlands. (2002).

20. Clemens Maidhof and Stefan Koelsch. Effects of selective attention on syntax processing in music and language. *Journal of Cognitive Neuroscience*. 23.9, 2252–2267. (2011).

21. Jessica Slater and Mina Kraus. The role of rhythm in perceiving speech in noise: a comparison of percussionists, vocalists and non-musicians. *Cognitive processing*. Springer. 17.1, 79–87. (2016).

22. Ferreri, Laura, and Laura Verga. Benefits of Music on Verbal Learning and Memory. *Music Perception: An Interdisciplinary Journal*. 34.2, 167–182. (2016).

23. Heinke, Dietmar, and Andreas Backhaus Modelling visual search with the selective attention for identification model (VS-SAIM): a novel explanation for visual search asymmetries *Cognitive computation*. 3.1, 185–205. (2011).

24. Tu, Z., Abel, A., Zhang, L., Luo, B., and Hussain A new spatio-temporal saliency-based video object segmentation. *Cognitive computation*. 8.4, 629–647. (2016).

25. Riche, N., Mancas, M., Culibrk, D., Crnojevic, V., Gosselin, B., Dutoit, T. Dynamic saliency models and human attention: a comparative study on videos *Asian Conference on Computer Vision. Springer, Berlin, Heidelberg*. 586–598. (2012).

26. Thornton W Burgess. *The Adventures of Reddy Fox*. 1–86, Courier Corporation, Chicago, IL, USA. (2012).

27. Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, Elsevier, Amsterdam, Netherlands. (2009).

28. Connie Mayfield and Sherry Moss. Effect of music tempo on task performance. *Psychological Reports*, 65(3 -Part 2):1283–90, SAGE Publications,

New York, NY, USA. (1989).

29. Letizia Marchegiani and Xenofon Fafoutis. A behavioral study on the effects of rock music on auditory attention. In *Proceedings of the International Workshop on Human Behavior Understanding*, 15–26, Springer, Berlin, Germany. (2013).

30. Hiyoshi-Taniguchi, Kazuko and Kawasaki, M and Yokota, Tatsuya and Bakardjian, Hovagim and Fukuyama, Hironori and Cichocki, Andrzej and Vialatte, François B EEG correlates of voice and face emotional judgments in the human brain. *Cognitive Computation*, 7.1:11–19, SAGE Publications, New York, NY, USA. (2015).

31. DeLiang Wang. On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines*, 181–197. Springer, Berlin, Germany. (2005).

32. DeLiang Wang, Ulrik Kjems, Michael S Pedersen, Jesper B Boldt, and Thomas Lunner. Speech intelligibility in background noise with ideal binary time-frequency masking. *The Journal of the Acoustical Society of America*, 125(4):2336–2347, Melville, NY, USA. (2009).

33. Richard Lyon. A computational model of filtering, detection, and compression in the cochlea. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 7:1282–1285, IEEE, New York, NY, USA. (1982).

34. Lagrange, Mathieu, Roland Badeau, and Gaël Richard. Robust similarity metrics between audio signals based on asymmetrical spectral envelope matching. *Acoustics Speech and Signal Processing (ICASSP), IEEE International Conference on. IEEE*, pp. 405–408. (2010).

35. Rabiner, Lawrence and Juang, Biing-Hwang. Fundamentals of speech recognition *PTR Prentice Hall* (1993).

36. Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proceeding of the International Society for Music Information Retrieval Conference (ISMIR)*, 255–266, Canada. (2010).

37. Yolanda Vazquez-Alvarez and Stephen A. Brewster. Eyes-free multitasking: the effect of cognitive load on mobile spatial audio interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, 2173–2176, ACM, New York, NY, USA. (2011).

5 Tables

Table 1 List of the songs used in the experiments along with their characteristics.

Song Code	**Artist**	**Song Title**	**Dynamics**	**Distortion**
ND	The National	Runaway	Low	No
D	Mother Love Bone	This Is Shangrila	Low	Yes
HD	The Pixies	Gouge Away	High	Both

Table 2 Precision and F1 score for all different kinds of masking conditions.

	Masker Type	Silence	ND	D	HD
Precision	Mean, μ	0.758	0.496	0.639	0.468
	Standard Deviation, σ	0.156	0.236	0.226	0.215
F1 Score	Mean, μ	0.822	0.581	0.692	0.547
	Standard Deviation, σ	0.131	0.236	0.192	0.199

Table 3 The response time of the subjects for all different kinds of masking conditions.

Masking Type	Silence	ND	D	HD
Mean, μ (s)	0.824	1.347	1.247	1.254
Standard Deviation, σ (s)	0.211	0.336	0.327	0.331

6 Figures

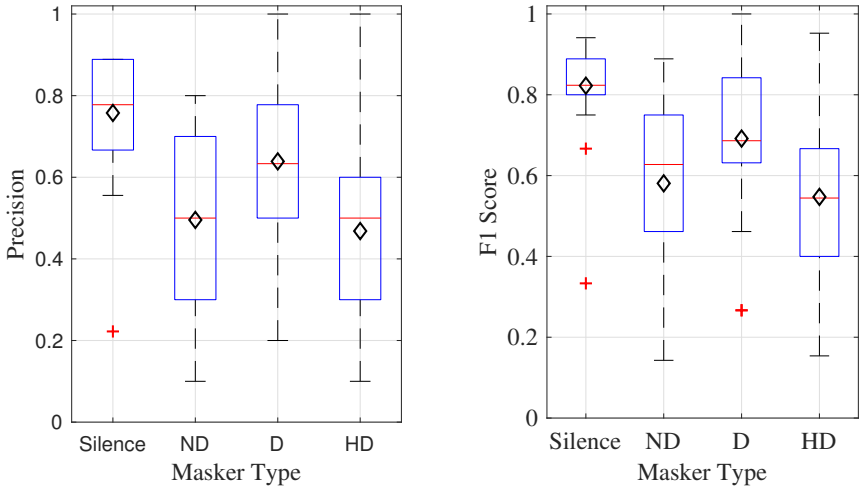

Fig. 1 Precision and F1 score for all different kinds of masking conditions. On each box, the central line indicates the median, the bottom edge of the box indicates the 25th percentile, and the top edge of the box indicates the 75th percentile. The whiskers extend to the maximum and minimum data points, excluding outliers ($\pm 2.7\sigma$) that are marked individually. The diamonds indicate the mean.

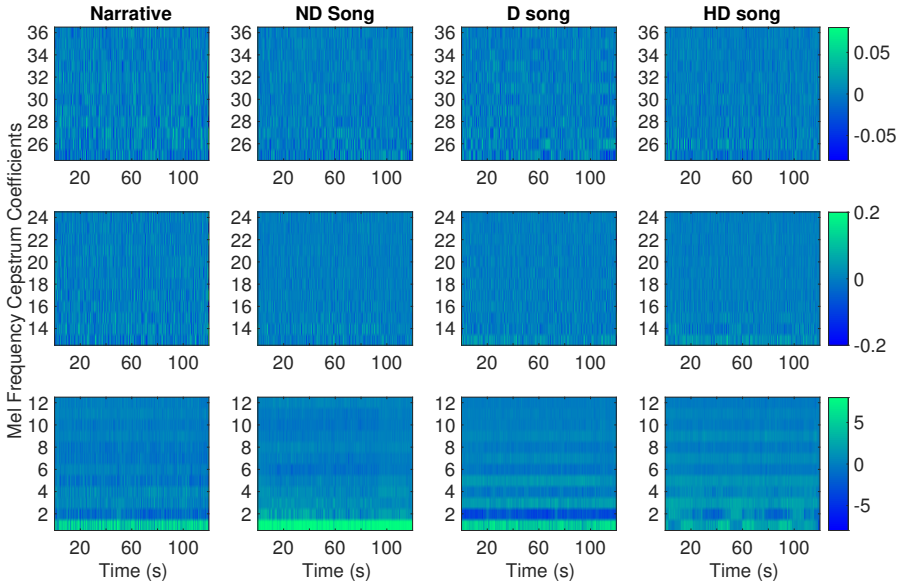

Fig. 2 Mel Frequency Cepstrum Coefficients (MFCCs) of different masker types: silence, songs without distortion effects (ND), songs with distortion effects (D), and songs characterised by high dynamics (HD). The figure plots 36 coefficients: the first 12 MFCCs, the 12 corresponding delta coefficients (delta MFCCs), and the 12 corresponding delta-delta coefficients (delta-delta MFCCs). All 36 coefficients are used to evaluate the similarity between the D and the ND songs; only the first 12 of those are used to analyse the dynamics of the songs.

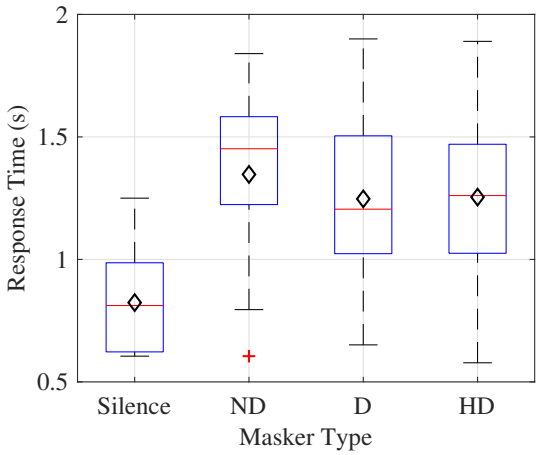

Fig. 3 The response time of the subjects for all different kinds of masking conditions. On each box, the central line indicates the median, the bottom edge of the box indicates the 25th percentile, and the top edge of the box indicates the 75th percentile. The whiskers extend to the maximum and minimum data points, excluding outliers ($\pm 2.7\sigma$) that are marked individually. The diamonds indicate the mean.